Influencia y aplicación de papeles sintácticos e información semántica en la resolución de la anáfora pronominal en español

Maximiliano Saiz Noeda
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
max@dlsi.ua.es

Resumen: Esta Tesis presenta un profundo estudio de la influencia y el uso de papeles sintácticos e información semántica en la resolución de la anáfora pronominal en español, en concreto la generada por pronombres personales, demostrativos, reflexivos y omitidos. Se plantea una combinación de ambas fuentes de conocimiento para proponer una base lingüística, científica y metodológica de resolución basada en información enriquecida que incorpora morfología, sintaxis y semántica.
Palabras clave: Resolución de la Anáfora, Papeles Sintácticos, Semántica, Ontología, WordNet

Abstract: This Thesis presents a deep study of the influence and the use of syntactic roles ans semantic information in the pronominal anaphora resolution for Spanish and, specifically, the generated by personal, demonstrative, reflexive and omitted pronouns. A combination of both knowledge sources is proposed in order to propose a linguistic, scientific and methodologic base basis based on enriched information that integrates morphology, syntax and semantic.
Keywords: Anaphora Resolution, Syntactic Roles, Semantic, Ontology, WordNet

1 Introducción

Esta Tesis Doctoral presentada el día 5 de junio de 2002 por Maximiliano Saiz Noeda para la obtención del título de Doctor Ingeniero en Informática ha sido desarrollada bajo la dirección conjunta del Dr. Manuel Palomar Sanz, de la Universidad de Alicante, y de la Dra. Lidia Moreno Boronat, de la Universidad Politécnica de Valencia. El Tribunal de Tesis, compuesto por los doctores Horacio Rodríguez Hontoria, Antonio Ferrández Rodríguez, Natividad Prieto Sáez, Ruslan Mitkov y Alfonso Ureña López concedió por unanimidad la calificación de Sobresaliente *Cum Laude*.

2 Motivación

La resolución de la anáfora ha sido durante las últimas dos décadas una preocupación de lingüistas e informáticos. Esta tarea, considerada por muchos como una de las más importantes dentro del tratamiento de la ambigüedad en el Procesamiento del Lenguaje Natural, ha sido abordada desde distintos puntos de vista por los sistemas más variados.
Algunos *métodos de conocimiento limitado*, han resuelto la anáfora con el uso de conocimiento pobre (sin análisis sintáctico) o a partir de análisis parciales o completos. Estos trabajos, realizados para el inglés (Hobbs, 1978; Lappin y Leass, 1994; Kennedy y Boguraev, 1994; Baldwin, 1997; Mitkov, 1998) y para el español (Ferrández, 1998; Palomar *et al*. 2001), coinciden (si bien no hacen uso de ella) en la necesidad de la semántica como fuente esencial para la correcta resolución de la anáfora.
Esta necesidad ha llevado a distintos autores a plantear *métodos de resolución enriquecidos* que si bien combinan la semántica y la sintaxis (Carbonell y Brown, 1988; Rich y Luperfoy, 1988; Kameyama, 1997; Mitkov, 1994; Carter, 1986), lo hacen para el inglés, en dominios muy restringidos o con definiciones puramente manuales de jerarquías y rasgos, lo que confiere un carácter más anecdótico al uso de información semántica en comparación con el resto de las fuentes de conocimiento.
Así mismo, otros *métodos alternativos*, incorporan los papeles sintácticos en patrones de co-ocurrencia a través de aproximaciones puramente estadísticas (Dagan e Itai, 1991).

Esta Tesis propone una base lingüística, científica y metodológica de resolución computacional de la anáfora pronominal en español (pronombres personales, demostrativos,

reflexivos y omitidos) que incorpora, de forma automática, información enriquecida con papeles sintácticos e información semántica.

3 Aportaciones de la Tesis

Las principales aportaciones de la Tesis son:

- *Contextualización y definición del fenómeno de la anáfora* que la relaciona con otros como la elipsis o la deixis y la clasifica de forma original según la naturaleza sintáctica del elemento anafórico entre otros criterios.
- *Exhaustiva revisión del estado del arte* basada en la clasificación de métodos expuesta anteriormente.
- *Estudio de las fuentes de conocimiento* que intervienen en el proceso de resolución de la anáfora y de recursos que las proporcionan.
- *Propuesta de un método de conocimiento limitado*, basado en un conjunto de restricciones y preferencias de carácter morfológico y sintáctico. Se presentan los resultados de su evaluación que se comparan con los obtenidos por otros métodos clásicos implementados y adaptados al español.
- *Propuesta de un etiquetado enriquecido* del corpus de entrada que cubra las necesidades de anotación que plantea el método enriquecido en lo referente a los papeles sintácticos y a los sentidos correctos de las palabras desde WordNet.
- *Propuesta del método enriquecido de resolución de la anáfora pronominal en español* (**ERA**) que incorpora las fuentes de conocimiento provenientes de los papeles sintácticos y la información semántica.
- *Construcción de un banco de pruebas* para la evaluación del método **ERA**, diseñado específicamente para determinar la influencia de las diferentes fuentes de conocimiento en el proceso de resolución de la anáfora.
- *Análisis de la influencia de las distintas fuentes de información* en la resolución de la anáfora con el método **ERA**.

Bibliografía

BALDWIN, BRECK (1997). "CogNIAC: high precision coreference with limited knowledge and linguistic resources", en *Proceedings of ACL/EACL workshop on Operational factors in practical, robust anaphor resolution*, págs. 38–45.

CARBONELL, JAIME G. Y RALPH D. BROWN (1988). "Anaphora resolution: a multi-strategy approach.", en *Proceedings of 12th International Conference on Computational Linguistics (COLING'88)*, págs. 96–101, Budapest, Hungary.

CARTER, DAVID M. (1986). *A shallow processing approach to anaphor resolution*, tesis doctoral, University of Cambridge.

DAGAN, IDO Y ALON ITAI (1991). "A statistical filter for resolving pronoun references", *Artificial Intelligence and Computer Vision*, págs. 125–135.

FERRÁNDEZ, ANTONIO (1998). *Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gramáticas de unificación de huecos*, tesis doctoral, Universidad de Alicante.

HOBBS, JERRY R. (1978). "Resolving pronoun references", *Lingua*, 44, 311–338.

KAMEYAMA, MEGUMI (1997). *Intrasentential Centering: A case study*, cap. Centering Theory in Discourse, págs. 89–112, Walker, M.; Joshi A. and Prince, E., eds., Clarendon, Oxford.

KENNEDY, CHRISTOPHER Y BRANIMIR BOGURAEV (1996). "Anaphora for everyone: pronominal anaphora resolution without a parser", en *Proceedings of 16th International Conference on Computational Linguistics*, vol. I, págs. 113–118.

LAPPIN, SHALOM Y HERBERT LEASS (1994). "An algorithm for pronominal anaphora resolution", *Computational Linguistics*, 20(4), 535–561.

MITKOV, RUSLAN (1994). "An integrated model for anaphora resolution", en *Proceedings of 15th International Conference. on Computational Linguistics (COLING'94)*, vol. III, págs. 1170–1176.

MITKOV, RUSLAN (1998). "Robust pronoun resolution with limited knowledge", en *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, págs. 869–875, Montreal (Canada).

PALOMAR, MANUEL, ANTONIO FERRÁNDEZ, LIDIA MORENO, PATRICIO MARTÍNEZ-BARCO, JESÚS PERAL, *MAXIMILIANO SAIZ-NOEDA* Y RAFAEL MUÑOZ (2001) "An Algorithm for Anaphora Resolution in Spanish Texts". *Computational Linguistics,* **27**(4), 545-567.

RICH, ELAINE Y SUSAN LUPERFOY (1998). "An Architecture for Anaphora Resolution", en *Proceedings of the 2nd Conference on Applied Natural Language Processing.*, 18–24, Austin, Texas.